ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

Apache Hadoop: Resourceful Big Data Management

Mr.NileshVishwasrao Patil, Mr.Tanvir Patel

ME Computer (I) Student, Vishwbharti Academy College of Engineering, Ahmednagar/ Pune University, Ahmednagar, Maharashtra, India.

ME Computer (I) Student, Vishwbharti Academy College of Engineering, Ahmednagar/ Pune University, Ahmednagar, Maharashtra, India

Abstract–Now days the growth of increasing data in one year is around double to existing data available up to previous year. Today's world is modern computer world, every public and private sector moving towards modern electronic world, also small data is moving towards big data. Hence there is need to distribute big data efficiently in distributed framework with replication for its importance. Big data is available in structured, unstructured and semi-structured data format. Relational database has fails to store this multi-structured data. Apache Hadoop is efficient, robust, reliable and scalable framework to store, process, transform and extract big data.Hadoop framework is open source and fee software which is available at ApacheSoftware Foundation. In this paper we will present Hadoop, HDFS, MapReduce and application projects to minimize efforts of developer to write MapReduce code.

Keywords: HDFS: Hadoop Distributed File System, Map Reduce, Hive, Name node, Data node, Task tracker, Job tracker, TB: Terabyte. Hive, Pig, HBase

I. INTRODUCTION

Todays the computer science enterprises is enduringmomentous changes since the development of new technologies, every public and private sector becomes computerized, increasing large amount electronic data termed: Big data etc.

Big data is data which is 1 TBor more than 1 TB. We can also defined big data is a data which cannot handle by single computer system. There is one more technical definition,data has large amount of volumes which comes from variety of sources with great velocity called as Big Data. Big data has three characteristics such Volume, Variety and Velocity so sometimes called 3 V's. Big data with its characteristics as shown in following fig. 1

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

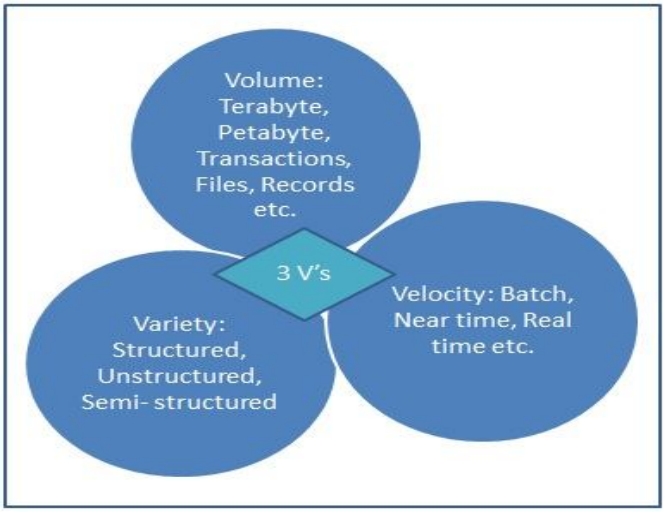

Fig. 1 3 V's of big data

Big data has generated from various sources such as web logs, traffic signal data, election data, world survey data, stock exchanges data, flight data, user transaction history, user interaction logs, RFID system generated files,social networking data like twitter, Facebook etc. with video and images. We must have to process and analyzethis big data therefore for analyze big data we will requiresa distributed framework as big data could not handle by single system since data is more than one terabyte. From last decades we have been using Message passing interface (MPI) as a distributed system to handle and process such huge volume of data.

Usually distributed systems has requires application server and storage area network (SAN). All data is available in SAN and all programs will run on application server. Before initiate the execution of program data moves towards application server from storage area network and after the execution of data application server writes the data on storage area network back. This distributed system has faces problems to storehuge amount of data robustly. Whenever a program execution start data is scale-up and when execution of program is finished then data is scale-down, it will affect huge dependency on network, travelling cost increases and scaling up and scaling down of large volume of data is not smooth process, also lots of processing power has consumed on transportation of big data.In typical distributed system partial failures are difficult to handle and which effect on our daily commercial business.One more problem is data synchronization is required during exchange of data. These are problem with typical distributed system to process big data. Therefore we requires new distributed framework which will remove above mentioned problems.New distributed system will efficiently store and process data with reliability. Apache Hadoop is distributed framework will removes all problem which faced by typical distributed system by providing efficient and reliable store and process of big data.

II. HADOOP FRAMEWORK

Today we just have surrounded by electronic data.As we have discussed above big data has generated from various sources like RFID system generated files, web logs, human interaction text, stock exchange data, social networking websites human interaction data, peoples upload video, images etc. Apache Hadoop is open source distributed framework to handle, extract, load, analyze and process big data. It facilitates to process large set of data across the cluster of computers using simple programming model. Before moving further we are going to see history behind Apache Hadoop. Google has published whitepaper on Google File System (GIS) and Map-reduce in 2004. Daug Cutting reads those papers, and he has extends Apache Nutch projects with the help of Google papers and developed Hadoop framework. He joined Yahoo in 2006 and great journey of Hadoop starts in Yahoo. And now Hadoopframework is Apache open source project.

Hadoop is open source framework for writing and running distributed applications that process large amount of data. Key

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

distinctions of Hadoop are Accessible, Robust, Scalable and Simple. Accessible: Apache Hadoop is runs on large cluster of commodity hardware, no need to purchase expensive hardware and also it runs on cloud computing. Robust: See Hadoopcluster consists of commodity hardware so possibility to occur failure but whenever failure occurs we can easily addressed failure to recover it.

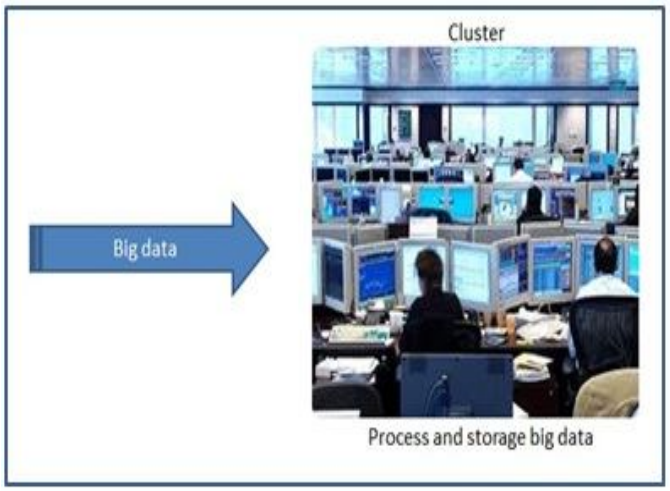

Fig. 2Storage and Process big data by cluster

Scalable: Framework scales linearly to handle big data by adding extra commodity nodes to cluster. Simple: It allows users to easily handle framework by using simple programming model [1]. How big data is store and process by cluster shown in above fig. 2

Now we are going to grasp how the transportation cost is minimized by Hadoop framework with example. Suppose we have to stores 100 TB data and to process this data require to write program using simple programming model but size of program not more than 10MB. In typical distributed system data is traveled to words the program and Apache Hadoop framework program is traveled towards data. Since the size of program is very small as compare to size of data. Thus typical distributed framework is transport 100TB data while Hadoop framework is transport 10MB data before execution of processing program. Above examples show how transportation cost is minimized by Hadoop distributed framework. Apache Hadoop cluster as shown in following fig. 3. In which multiple client can store and process data into cluster simultaneously.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

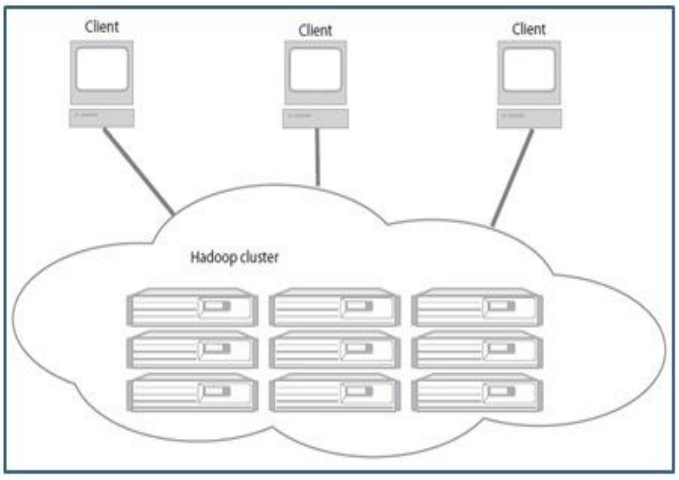

Fig. 3Many Parallel client store and process big data in Hadoop cluster [1]

Apache Hadoop is consistingof two main components: HDFS and MapReduce. Hadoop cluster consist of multiple machine in which running HDFS and MapReduce. Each machine in cluster is called as node.

Hadoop cluster has two types of nodes: Master and slave node, only one master and multiple slave nodes are available incluster. Hadoop is uses HDFS to storage and MapReduce to process the data.

III. HADOOPCLUSTER PROCESSING

Apache Hadoop framework consists of five types of daemons:Namenode, Datanode, TaskTracker, JobTracker, and Secondary Namenode. Namenode is stores all metadata information about cluster, like where data has stored, replication of data etc. It is running in master node of cluster. Namenode is take care how data block is broken down into multiple blocks with maintaining of replication of blocks. Secondary Namenode is replication of Namenode, if Namenode will crashat that time we can take backup from secondayNamenode manually to persist state of Namenode as before crash.

Actions → / Nodes ↓	Storage Data (HDFS)	Process Data (MapReduce)
Master Node (Only one)	Namenode (Extra: Secondary Namenode – For backup if any failure in Namenode)	JobTracker
Slave Node (Multiple)	DataNode	TaskTracker

Fig. 4 Jobs of each daemon in cluster

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

Datanode is available on each slave machine. Its take care the job of HDFS for each slave for storage data blocks. JobTracker daemon is take care about master node to process data. It has assigned task to Tasktracker using MapReduce programming model. TaskTracker is available on each slave to process data. As shown in following fig. 4 the jobs of each domain.

Now we are going to show how data is store and process by Hadoop with example and fig 5 as follows.Suppose we have 192 MB data and want to store this data in Hadoop cluster with two replication factor.

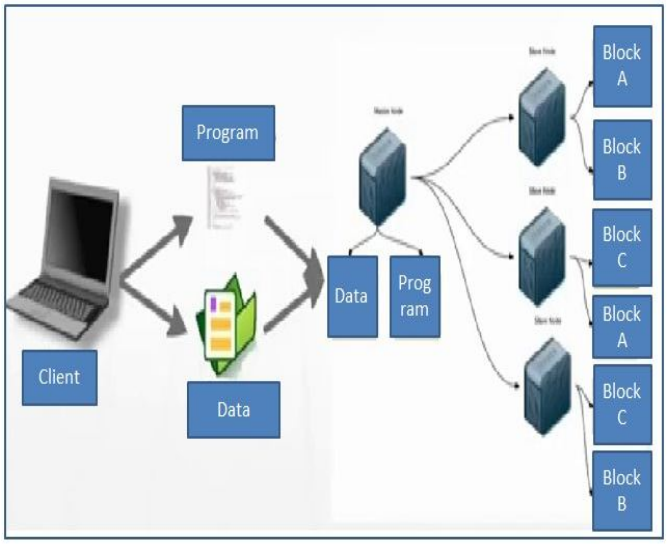

Fig. 5 Storage and process of data in cluster

In cluster before processing, file is broken down into blocks of size 64MB or 128MB and then moves block to different nodes. Then Hadoop framework will runs program to process blocks. JobTraker then scheduling program to each node and TaskTrackerwill process data. After complete storage and process of data the output is written back.

In our case 192MB data, we will divide that data into three blocks of size 64MB (64Mb * 3= 192MB). We want to store these three blocks with replication factor two for possibility to minimize data loss if any system fails and also for fast access. Three slave nodes are available in our cluster, so Hadoop frameworkplaces blocks, preferably put replicated block on different machine. As shown in fig 5.

Block A => stores at slave 1 and slave 2
Block B => stores at slave 1 and slave 3
Block C => stores at slave 2 and slave 3

Above all discussed things is done by Hadoop cluster itself, client is justprovidingdata in file and replication factor (Number times to replicate data).

IV. HADOOPDISTRIBUTED FILE SYSTEM

The HDFS is one of the core components of Hadoopand which is storage layer of Hadoopframework. Hadoop has its own file system based on GIS called as HDFS. HDFS is java based file system which can store large amount of structured, unstructured and semi-structured data. It is distributed, reliable, scalable, and fault tolerance file system. A typical file on HDFS is GB to TB or PT. HDFS architecture is shown in fig. 6 as follows.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

The HDFS file system is provides strong aggregate feature to blocks of data. When data is comes to master node in file then it will divides into multiple blocks for storing in cluster of nodes termed as: Fan-out and when client request for data then aggregate/collect blocks of data from cluster of nodes termed as: Fan-In. So that we can say HDFS has strong aggregate characteristics or it is based on aggregate design pattern.

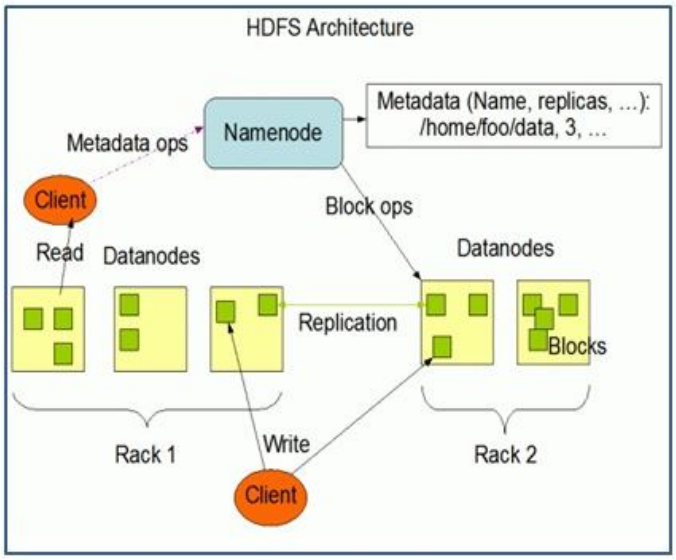

Fig. 6 HDFS architecture [4]

The HDFS is master/slave architecture which consist of one Namenode (for master) and multiple Datanodes (each per slaves) which shown in fig 6. Namenode contain metadata information and manages system namespaces. It also controls access to files by clients. The Namenode is executes file system namespace operations such as opening, renaming, closing the files stores in HDFS [2]. The file is divided into multiple blocks and those blocks are stored at different Datanodes.

Namespace ID of Namenode and Datanode must be same. If any incompatibility betweenNamenode and Datanodenamespace ID in cluster,we will getthe exception like java.io.IOException: Incompatible namespace ID.There are two ways to remove this incompatibility, first isto reformat Namenode but this is not good option and second by manually change namespace ID of slave to namespace ID of master. Namespace ID available to following directory:

/application/hadoopdata/temp: Directory created by user for HDFS.

For Namenode namespace ID at
/application/hadoopdata/temp/dfs/name/current/VERSION

And for Datanode namespace ID at
/application/hadoopdata/temp/dfs/data/current/VERSION

To remove this just copy the namespace ID from Namenode VERSION file to Datanode VERSION file which is available in above directory.

HDFS is operates on top of native UNIX file system on cheap commodity hardware and also performs data replication job. It is read only file system and random writes does not allowed. Namenode is up all time because it belongs to master node of cluster. The HDFS file system mostly developed for Batch processing instead of interactive use by users. Since HDFS is performingits task efficiently when files are contains big data minimum 1TB, so that it is mostly prefer for Batch

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

processing. HDFS file system is operating system independent so it can be run on heterogeneous operating system. HDFS file system is called "Write-Once, Read-many" because could not change data once push into cluster of nodes.

V. MAP REDUCE

MapReduce is data processing component of Hadoop framework and it is compute/process layer of Hadoop like HDFS is storage layer. It is programming model based on Java programming. Process layer is consisting of two phases: one is Map and second is Reduce. There is one more layer between these two phasescalled as Sort and Shuffle. JobTracker (for Namenode) and TaskTracker (perDatanodes) take care of MapReduce job. Logical architecture of Hadoop is shown in fig.7 as follows. It is also called as MapReduce process.

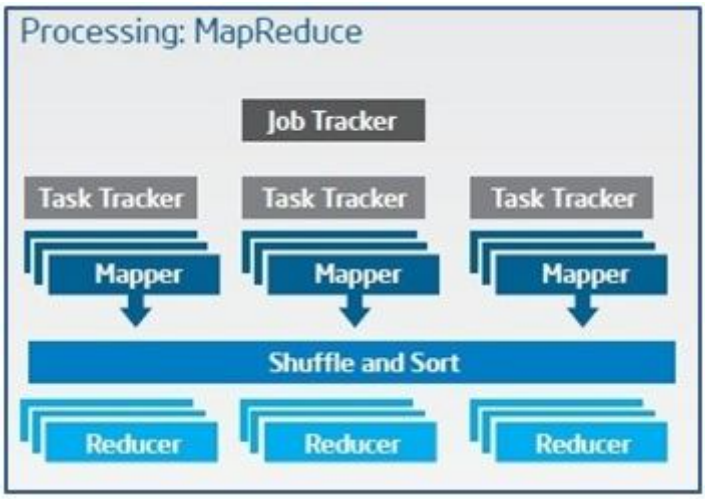

Fig. 7 MapReduceProcess[3]

MapReduce programing model is similar like programming languages (C, C++, C#, and Java)but it is difficult to understand and write programs. Therefore application projects have introduced to minimize efforts for writing MapReduce code. There are list of application projects available such as "Hive, Pig, HBase, Flume, Oozie, Ambari, Avro, Mahout, Sqoop, HCatlog, BigTop" etc.

Hive was developed by Facebook and now it available open source.It is data processing structure based on Hadoop which runs top of Hadoopframework. Hive application project is allows developers to write job of processing data in queries like SQL language. HiveQL is a language provided by Hive. It minimizes efforts of programmer to write MapReduce job. Hive application project is converts our HiveQL query into MapReduce program.

Pig application project was developed by Yahoo based on Haddop which also runs on top of Hadoop framework equivalent to Hive. Pig Latin language is used by Pig, which is easier to write data processing job. Pig application project is converts Pig Latin to MapReduce program and perform desired task without writing MapReduce program by developer.HBase is non-relational database that allows for low-latency, quick response in Hadoop. It supports transactional capabilities to Hadoop framework which allows users to behavior updates, inserts and deletes. Facebook uses mostly HBase on top of Hadoop. The list of application projects is increasing day by day.

The process flow of MapReduce job is show in following fig. 8. MapReduce is based on key-value pair idea. As shown in fig 8, the input data set is divided into 'n' splits (divide and conquers method).

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

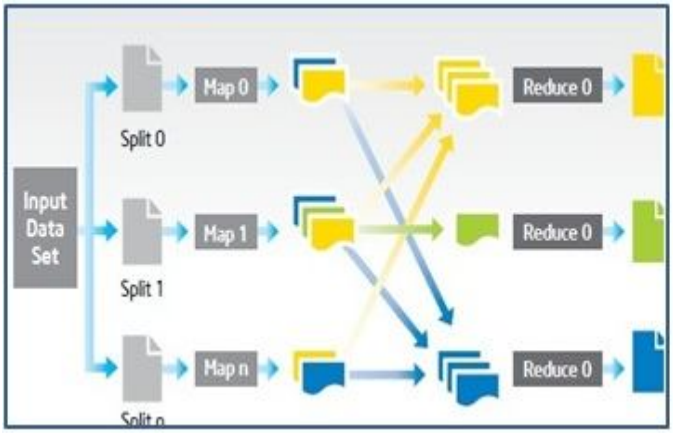

Fig. 8 Process flow[3]

Then 'n' map is performed functionality for each split, formerly sort and shuffle task perform before reducing, and finally the aggregating the result of data in Reduce phase.

Hadoop Users: "Amazon/A9, Facebook, IBM, Google, Yahoo, New York Times, Fox Interactive media and so on".
Major contributors: "Apache, Yahoo and Cloudera".

VI. CONCLUSION

In this paper we haveopenedthe role of Hadoop framework in big data.Apache Hadoop is designed to distribute large volume of structured, unstructured and semi-structured data across nodes in cluster with commodity hardware.In thispaper we have also discussed Hodoop Distributed File System, MapReaduce, daemons of Hadoop (NameNode, DataNode, TaskTracker, JobTracker and Secondary NameNode)and application projects for Hadoop to minimize the efforts of developer for writingMapReduce program. We conclude that Apache Hadoop is efficient, robust, reliable and scalable framework to store, process, transform and extract big data in cluster of nodes.

ACKNOWLEDGMENT

This paper is completed only because support from each and every one including: Government Polytechnic, Ahmednagar, teachers, colleague, parents, friends and my students.

Especially, my acknowledgment of gratitude toward the following important persons:

First I would like to thankMr. M. Kshirsagar Sir,Mr. Prabhudev sir, Mr. Natikar sir, and Mr. Jaypal sir,and my classmates to their support and encouragement.Second,

I sincerely thank to my parents andS. A. Bhalerao who provide the advice and financial support.

Last but not least, my late grandfather and late grandmother for their love. This research paper will not be possible without all of them.

REFERENCES

[1] Chuck Lam, "Hadoop IN ACTION" in Manning Publication Co., Stamford CT, USA, 2011, pp. 1-173.
[2] Vidyasagar S. D. (2013). Role of Hadoop in Information Technology era. Global Research Analysis [Online]. 2(2), pp. 100-101. Available Vin Sharma, "Extract, Transform, and Load Big Data with Apache Hadoop" whitepaper [Online], Intel Corporation, 2013. Available
[3] Michael G. Noll. *Running Hadoop on Ubuntu Linux (Single Node Cluster)* [Online]. Available: Website links: Apache Hadoop

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 4, April 2014

Two days National Conference – VISHWATECH 2014

On 21st & 22nd February, Organized by

Department of CIVIL, CE, ETC, MECHNICAL, MECHNICAL SAND, IT Engg. Of Vishwabharati Academy's College of engineering, Ahmednagar, Maharastra, India

BIOGRAPHY

Author[1]: Mr. Nilesh V. Patilis pursuing ME Computer from Vishwbharti Academy College of Engineering Ahmednagar. He has more than four years of experience involving around two years of industrial experience as Software engineer. He has been working as System Analyst in Government Polytechnic, Ahmednagarsince 24thNov 2011.

Author[2]: Mr. Tanvir Patelis pursuing ME Computer from Vishwbharti Academy College of Engineering Ahmednagar.